Virtual Institute of Microbial Stress and Survival
DOE Genomes To Life Project
Progress Report: October, 2003

I. Overview

The objective of this monthly progress report is to provide an update of the technical and administrative actions from the previous month as well as forecast upcoming progress for the VIMSS Genomes to Life Project at Lawrence Berkeley National Laboratory. I want to remind everyone how important to make sure everyone is communicating. The discussion boards (http://genomics.lbl.gov/~aparkin/discus) provide a forum for people to ask questions about direction of the project, priorities, and technical issues that can be read and answered by the entire group. I know email is often the most efficient means but it does privatize some of the important communications. Also, posting project data and information to BioFiles (https://tayma.lbl.gov/perl/biofiles) is EXTREMELY important. We are in the process of adding user help files to BioFiles – if you have user questions, please contact Keith Keller (tel: 510.495.2766 or email: kkeller@lbl.gov). This is the best metric I can give to the DOE leadership that we are making progress aside from the VIMSS website. Please make us and yourselves visible by donating data and information to the website.

II. Applied Environmental Microbiology Core

LBNL

SR-FTIR. We are setting up HPLC and GC/MS systems for exopolysaccaharide (EPS) analysis. This is important for the project because this would help to establish the SR-FTIR method, and to complement the metabolite analysis in Keasling's lab. Although EPS is a family of bacterial metabolites which plays an important role in bacterial responses to stress factors, EPS is not currently being considered in Kealsing's effort. We are also preparing the manuscript "A Direct observation of *Desulfovibiro vulgaris* response to sudden influx of oxygen" for submission.

Biomass Production. Additional biomass deliveries were made this month from the O_2 stress experiments.

Experiment 5: time course samples to ORNL
Experiment 6: time course samples to ORNL

It was determined after the first shipment that 50 ml of sample volume was insufficient, so the second sampling was increased to 300 ml per sample. Past biomass production events have been mostly successful and many lessons have been learned. For future production, we will tailor sampling and biomass amount to specific needs of the group requesting. No further biomass production is currently scheduled, but we are capable of shipping new cells within one week of a request.

Our QA/QC purity checks for biomass production will soon include an antibody screen for the identification of *D. vulgaris*. We have begun the preparation of *Desulfovibrio vulgaris* antiserum for both O and H antigens. Pacific Immunology will produce the antibody which will be IgG purified and conjugated to FITC for visualization of bound antibody using epifluorescence microscopy. The protocol used will be posted on Biofiles.

For future biomass productions, we are considering filtering the samples as a viable alternative to centrifuging. We purchased from Millipore two 47 mm stainless filter holders and two packs of 0.45 µm LCR hydrophilic PTFE membrane filters. After preliminary filtration tests, it was discovered that the 0.45um filter clogs easily and that prefiltering or using a larger filter holder will be necessary. More filtration tests will be conducted to determine the best prefiltering membrane filter to use and how much culture can be filtered at various optical densities ranging from 0.3 to 0.6 OD.

Other progress this month includes development of growth curve assay using a 96 well plate. This method allows for simultaneous and automatic measurement of 10 growth conditions with six replicates per condition. The plate is prepared in the anaerobic chamber and sealed with a silicone sealant. The plate is incubated in the plate reader at 30°C. We purchased new software for the reader that allows us to automatically log the OD of the plates for several days. The table below summarizes the conditions that we have looked at in the plate thus far. A sample growth curve of *D. vulgaris* in standard LS4D is also shown.

In November, we will be having a 2-day FairMenTec bioreactor collaboration at our laboratory with representatives from David Stahl lab (Beto Zuniga) and Judy Wahl's lab (Bill). The company that built the system is unable to send a representative for installation and training, so this collaboration will be to help our laboratory set up the bioreactors quickly and to set up a protocol that builds on the experience of these other two laboratories. Once functional, these bioreactors will eventually augment or replace the biomass production in bottles.

University of Washington

A pilot fixed bed biofilm reactor was inoculated with *D. vulgaris* and operated for 44 days before termination. At termination the reactor was consuming all lactate in a 16mM lactate feed. The influent rate was 0.5ml/min and recirculation flow rate was 50ml/min. The volume of the vessel is 600 ml, the estimated void volume of the vessel and loop is approximately 300ml where the vessel void volume being 230 ml. The 3 mm diameter glass beads which used to fill the reactor have a volume of 0.014cm^3 each, giving an estimated 26428.5 beads in the vessel. Nine subsamples of approximately 265 beads each were taken for analysis at run termination. Biofilm was dissociated from the beads by vortexing and recovered cells quantified by direct counting using a Petroff–Housser chamber. These measurements suggested that approximately 1.6x10^6 cells had colonized each bead. The number of cells per bead on day 18 was comparable to day 44, indicating that a steady state was achieved (decay/detachment balanced by growth/attachment). In it current configuration the total dry biomass on the beads in the system

was estimated to be between 100 – 200 mg ($5x10^{10}$cells). We anticipate that this will be sufficient for transcriptome and proteome analyses.

Growth of *D. vulgaris* was evaluated in the FairMenTec Bioreactor in chemostat mode using the LS4D medium supplemented with lactate (60 mM) and sulfate (50 mM).

$TiCl_3$-citrate was added at 1/3 of the recommended concentration (0. 117g/L) was added as reducing agent. The culture was initiated in batch mode and switched to chemostat mode at a flow rate of $0.075h^{-1}$ when biomass reached an OD 600nm of 0.3 (fig. 1.). At 20 hours operation the optical density fell to 0.2 and the flow rate reduced to $0.05h^{-1}$ to prevent wash out. The optical density stabilized at 1.0 OD_{600} and the flow rate was again increased to 0.075 (at this time we also observed formation of wall growth). For the final 300 hours of operation the reactor was run under conditions of sulfate limitation (60 mM lactate and 25 mM sulfate in the feed). During this final stage of operation the cell density gradually decreased to 0.6 OD_{600}, with a residual 52 mM lactate in the medium indicating that sulfate limitation was not achieved.

We continued the isolation of sulfate reducing strains from FRC area 2 enrichments. We previously reported isolation of colonies in agar amended with lactate, acetate or hydrogen and carbon dioxide. These were transferred to liquid medium amended with same corresponding electron donors.

To further characterize *D. vulgaris* strains isolated from Lake DePue, chromosomal DNA of two of them (DP4 and DP8) was digested with EcoRI, NotI, SwaI, Pst , SmaI and I-CeuI and analyzed by both conventional and by pulse field electrophoresis. Differences in pattern of fragments were observed in several cases demonstrating that the two Lake DePue isolates are very similar but not identical to *D. vulgaris* Hildenborough.

Mass spectrometer QIC20 (Hiden Analytical Inc.) for gas analysis was received and installed.

Immediate future work

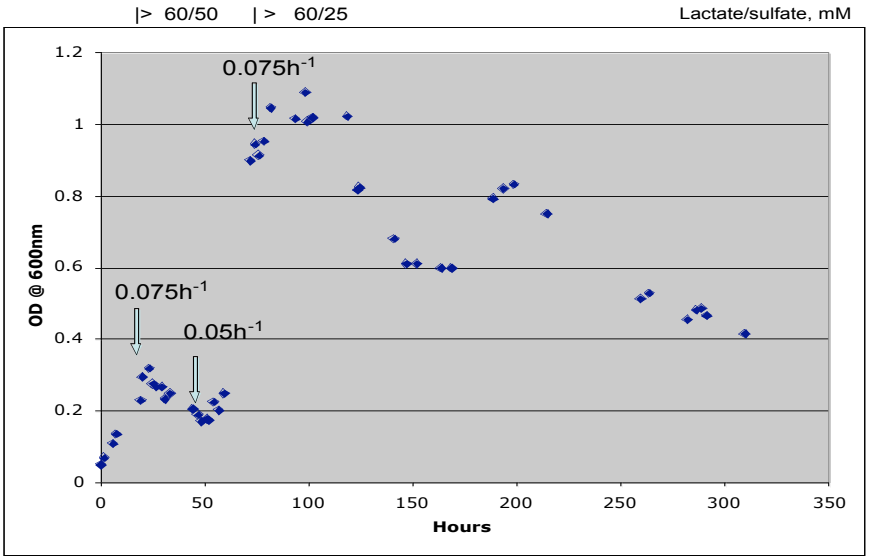
Continue analysis of the physiology of syntrophic association between each of two SRB strains and methanogen

- Determine stoichiometry of substrates and metabolites (lactate consumption, and hydrogen, acetate, methane production)
- Quantify cell numbers and biomass of methanogen and sulfate reducers.

Continue isolation and characterization of sulfate reducing bacteria from FRC enrichments. Amplify and sequence 16S rRNA and *dsrAB* genes from FRC isolates.

The new fixed bed reactor will be equipped with three daughter reactors and inoculated with a pure culture *D. vulgaris* Hildenborough. Biofilm development (number of cells on a bead, thickness of biofilm) and physiology (substrate consumption rate) will be monitored. Cells will be collected for providing material for transcriptome and proteome analyses.

Figure 1. Chemostat Growth of *D. vulgaris*



Diversa

Progress

- Carl and Denise visited LBNL to perform large and small insert DNA extractions from the following ^{238}U contaminated soil samples:

  Area 1: FB060-01-00, 0-25 inches, (2-20-03)
  Area 2: FB052-01-00, near DP06, 20-25 feet, top, (2-18-03), ~200 ppm ^{238}U.
  Area 3: FB056-01-34, near FW009, (2-19-03), ~25 ppm ^{238}U.

- The five samples described below have been amplified to obtain enough DNA for library construction. The amplified DNA has also been analyzed using diversity indexing to determine the complexity.

Genomes To Life Environmental Samples (October)

Sample ID	Large Fragment DNA Extraction	sample name	MDA	Diversity Indexing	DI Results	Small Insert Library	complete	Large Insert Library	complete
Area1 FB060-01-00	50g soil 10/1	1aL	10/8	10/17		10/10		10/10	
Area2 FB052-01-00	50g soil 10/1	2aL	10/8	10/17		10/10		10/10	
Area3 FB056-01-34	50g soil 10/1	3aL	10/8	10/17		10/10		10/10	
Area2 FB052-01-00	50g soil 10/2	2bL	10/8	10/17		10/10		10/10	
Area3 FB056-01-34	50g soil 10/2	3bL	10/8	10/17		10/10		10/10	

Issues

- Screening of libraries is dependent on selection of a suitable target.

Actions

- Newly extracted DNA from samples collected at Areas 1, 2 and 3 is currently being used for large and small insert library construction.

III. Functional Genomics Core

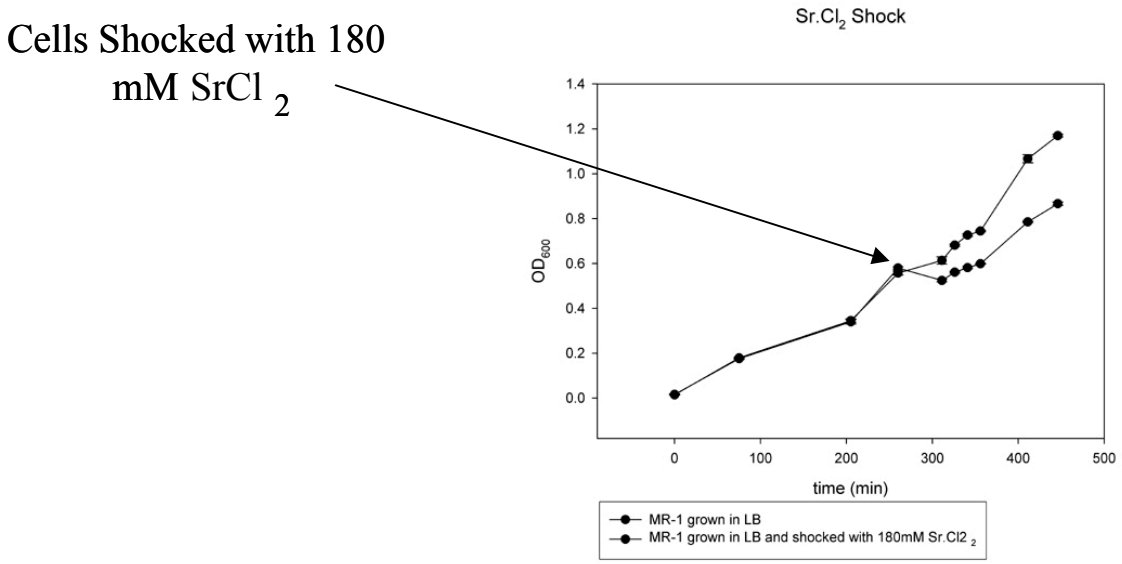
Oak Ridge National Laboratory (Transcriptomics)

Progress since last report

Shewanella

- pH stress datasets were sent to Sarah Wang in Adam's group for further analyses.

- We are continuing with microarray studies of the response to *Shewanella* to the heavy metal strontium. A time-series experiment was completed that examined the temporal gene response to 180 mM strontium at 7, 30, 60, and 90 minutes.
- A number of iron transport genes were highly induced in response to high levels of strontium, including open reading frames (ORFs) predicted to encode siderophore biosynthetic proteins.



Gene ID	Time (min)				Gene	Gene Product
	t=7.5	t=30	t=60	t=90		
SO3669	1.282	46.142	177.354	168.03	*hugA*	heme transport protein
SO3670	6.161	83.816	155.477	225.49	*tonB1*	TonB1 protein
SO3671	3.306	197.4	337.067	621.96	*exbB1*	TonB system transport protein ExbB1
SO3672	2.441	64.688	126.004	173.16	*exbD1*	TonB system transport protein ExbD1
SO3673	1.519	37.288	87.683	82.232	*hmuT*	hemin ABC transporter, periplasmic hemin-binding protein
SO3674	1.5	27.556	60.378	170.403	*hmuU*	hemin ABC transporter, permease protein
SO3675	1.176	42.675	115.959	111.29	*hmuV*	hemin ABC transporter, ATP-binding protein

SO3030	1.318	20.783	32.875	49.2747		siderophore biosynthesis protein	
SO3031	0.964	9.2082	134.9	245.771		siderophore biosynthesis protein, putative	
SO3032	0.694	4.9385	182.35	126.924		siderophore biosynthesis protein, putative	
SO3033	0.617	4.1357	88.071	89.9729		ferric alcaligin siderophore receptor	
SO3034	2.189	154.41	32.012	35.8445		ferric iron reductase protein, putative	
SO1937	1.2	1.6	2.2	1.9	fur	ferric uptake regulation protein	Regulatory functions	DNA interactions

Desulfovibrio

- RNA extraction and microarray hybridization with LBNL cell pellets (Terry's group) from 50-ml culture resulted in insufficient total RNA, ranging from 5 □g to 8 □g. Subsequently, microarray experiments were not successful.

- RNA extraction and microarray with LBNL cell pellets from 300-ml culture yielded around 30 μg of total RNA, barely enough for 3 replicate microarray hybridizations. However, RNA concentration was low. As a result, vacuum dry and re-suspension were necessary prior to the hybridization protocol. Four samples (2 for control and 2 for treatment) were shipped and so used only for testing purposes. Quality of the microarray images was okay.
- Microarray trials using gDNA as reference are in progress.
- Salt shock (5 time points): RNA has been extracted and verified by gel analysis.
- Trial for nitrite shock (6 time points): hybridizations have been completed and imaging is in process.

Geobacter

- Construction of *Geobacter* arrays are in progress.

Diversa (Proteomics)

Objectives

- Proteomics analysis of *D. vulgaris* upon O_2 stress-response

Progress since last report

A. Comprehensive 3D LC-MS/MS analysis of *D. vulgaris* O_2-stressed samples

3D LC MS/MS of following 8 samples

Digested Sample	Protein Extract (mg)	Conc (ug/ul)	Digested Amount (ug)	Vol (ul)	Peptide samples
E3T0C1(X)a	14 from 580	2.44	1000	410	Rg whole cell from D vulgaris E3T0C1
E3T0C1(X)b					Pellets after acid treated E3T0C1(X)
E3T0V1(X)a	15 from 440	2.63	1000	380	Rg whole cell from D vulgaris E3T0V1
E3T0V1(X)b					Pellets after acid treated E3T0V1(X)
E3T1C1(X)a	12 from 490	2.13	1000	470	Rg whole cell from D vulgaris E3T1C1
E3T1C1(X)b					Pellets after acid treated E3T1C1(X)
E3T1V1(X)a	7 from 580	1.39	1000	720	Rg whole cell from D vulgaris E3T1V1
E3T1V1(X)b					Pellets after acid treated E3T1V1(X)

- 1X complete 3D analysis and database searching of above samples

Digested Sample	Sample name	Analysis time	Dexter searched results(protein IDs)	Total identifications
E3T0C1(X)	E3T0C1500ugRgS100603Tinman	5 days	1176	1270
E3T0C1(X)	E3T0C1halfRgS100603Tiger	3 days	808	
E3T0V1(X)	E3T0V1500ugRgS103103Tinman	5 days	1167	1251
E3T0V1(X)	E3V1T0Rgphalf102903animal	2 days	776	
E3T1C1(X)	E3T1C1500ugRgS101303Tinman	5 days	1237	1429
E3T1C1(X)	E3T1C1RGPhalf101303Tiger	3 days	1114	
E3T1V1(X)	E3T1V1500ugRgS102103Tinman	5 days	1122	1295
E3T1V1(X)	E3T1V1RGPhalf101703Tiger	3 days	967	

B. ICAT analysis of D. *vulgaris* O2-stressed samples

Optimization of the ICAT LCMS method.

Variable tested: Quantity of proteome lysate that can be used in the ICAT protocol. 250, 500 and 1000ug protein from both E1T1C-1 and E1T1V-1 lysates (prepared as described in previous report) were used. To reduce disulfide bonds, 2ul TCEP solution (ICAT Kit, Applied Biosystems) was used per 50ug protein in sample. The proteome was denatured by boiling for 10minutes. Each E1T1C-1 sample was added to 1 vial of Light reagent. Each E1T1V-1 samples was added to 1 vial of Heavy reagent. Labeling proceeded for 4hours at 37^{o}C. The labeled samples were pooled, digested overnight with Trypsin. The pools of heavy and light peptides were purified using a CE (cation exchange, ICAT Kit) column and the resulting mix was passed through an affinity column (ICAT Kit). The eluants, consisting primarily of cysteine-labeled peptides, were treated with the cleavage reagent (ICAT Kit) to cleave off the biotin group. The final heavy and light tagged peptides were dried in a speed-vac and kept at –20^{o}C for MS runs.

The ICAT samples (ICAT250, ICAT500, ICAT1000) were analyzed using 1D LCMSMS:

Each ICAT sample was suspended in 10ul 0.1% HCOOH. 6ul of this was loaded onto the LC column (75micron, 300^{o}A, 20cm) using full loop pickup mode (Famos Autosampler, ~ 1/3rd actually loaded). A gradient of 2-30% of solvent A (5% ACN in 0.1% HCOOH) to B (80% ACN in 0.1% HCOOH) over a period of 1hour, was used to resolve the peptides. For every TOF spectra, MSMS data was obtained for 2 ions of +2 to +4 charge. Cut off was set at 100counts. *Unfortunately, this experiment cannot differentiate between the limiting factor being the quantity of labeling reagent, binding capacity of the affinity column or the detection limit of the QSTAR TOF MS instrument.*

The data in the 3 cases were analyzed using ProID and ProICAT (Applied Biosystems). The TIGR annotated *D. vulgaris* ORF FASTA file (GDV.pep) from the biofiles website was used to create interrogator databases for both ProID and ProICAT. Cycles 1000-3300 were used for each TIC.

H:L ratio => O2 stressed proteome at T5 vs control

ProID can identify proteins without the ICAT label and hence the difference between the number of hits in ProID vs ProICAT provides some idea as to how many non-ICAT labeled peptides leaked through the affinity column. It appears that using more proteins at the start of the procedure, improves ProICAT:ProID ratios especially at 90% confidence. A synopsis of the data is as follows:

Proteins found	**ProID**			**ProICAT**		
Confidence	50%	75%	90%	50%	75%	90%
ICAT250	106	79	33	42	9	5
ICAT500	136	89	29	58	13	6
ICAT1000	216	137	45	90	23	13

Peptides found	**ProID**			**ProICAT**		
Confidence	50%	75%	90%	50%	75%	90%
ICAT250	155	116	51	52	14	8
ICAT500	219	153	62	95	32	14
ICAT1000	319	205	80	117	36	25

Both ProID and ProICAT identified several proteins predicted to be affected in an O_2 stress. Shown below is a list of proteins ID-ed with > 90% confidence using ProICAT for the ICAT1000 TIC. Of these, proteins predicted by Judy and Sergey are highlighted.

Accession #	on	Peptides found
ORF01645	Protein	11
ORF00338	reductase, alpha subunit	10
ORF00303	translocase, SecA subunit (secA)	6
ORF05313	sulfite reductase alpha	4
ORF03227	bifunctional protein, putative	3
ORF04271	Rubredoxin	3
ORF01531	ferredoxin oxidoreductase, beta subunit, putative	2
ORF01685	class I (aspC4)	2
ORF03581	reductase, dissimilatory-type gamma subunit	2
ORF05598	domain, fis-type protein (b0502)	2
ORF01089	elongation factor G (fusA)	1
ORF03747	Protein	1
ORF05010	Protein	1
ORF05049	phosphate dikinase, PEP	1
ORF05307	acid synthase, putative	1
ORF01327	hypothetical protein	1
ORF01214	reductoisomerase (ilvC)	1
ORF02147	kinase (adk)	1
ORF01093	protein L3 (rplC)	1
ORF02413	pyrophosphokinase (relA)	1
ORF01032	Protein	1
ORF00875	Protein	1
ORF00824	hypothetical protein	1
ORF00778	Protein	1
ORF00315	hypothetical protein TIGR00096	1
ORFA00049	(EC1.11.1	1
ORF01122	protein L15 (rplO)	1
ORF03980	domain S-box protein	1
ORF04606	protein hydh (hydH)	1
ORF04575	formyltransferase (fmt)	1
ORF04462	Synthase, putative	1
ORF04301	Protein	1
ORF04950	ArCR, COG2043 superfamily	1
ORF04106	fatty acid transport protein, putative	1
ORF02121	large chain precursor (hybC)	1
ORF03990	oxidase, subunit GlcD (glcD)	1
ORF04735	system protein	1
ORF03929	transferase, group 2 family protein domain protein	1
ORF03818	protein L1 (rplA)	1
ORFA00114	Protein	1
ORF04961	Protein	1
ORF00018	U32 family	1
ORF02711	AhpC	1
ORF02453	ACR, COG1433 family	1
ORF04028	dehydrogenase (asd)	1

Candidates also found in the ProICAT analysis of the TICs (ICAT250, ICAT 500 and ICAT1000) with >90% confidence, are listed below.

Accession #	Annotation	Peptides Found
ORF00338	reductase, alpha subunit	8
ORF05313	sulfite reductase alpha	4
ORF04271	Unknown	2
ORF01081	adenylyltransferase (sat)	1
ORF01093	protein L3 (rplC)	1
ORF01214	reductoisomerase (ilvC)	1
ORF01685	class I (aspC4)	1
ORF01911	Protein	1
ORF02147	kinase (adk)	1
ORF02453	ACR, COG1433 family	1
ORF03581	reductase, dissimilatory-type gamma subunit	1
ORF03818	protein L1 (rplA)	1
ORF04462	synthase, putative	1
ORF04950	ArCR, COG2043 superfamily	1

ProICAT also provides corresponding Heavy:Light ratios. However, while several common proteins were found in the ProICAT analyses of the three data sets, their H:L ratios from the ICAT250, ICAT500 and ICAT1000 data do not agree with each other (See Raw data files, Dv-O2stress-ProID.xls and Dv-O2stress-ProICAT.xls). The same proteome lysate was used in all 3 labeling experiments. Labeling and MS data acquisition for all 3 were done together. Furthermore, the proteins were IDed with high confidence in 3 data sets by two different types of software.

Future work

- Further data analysis
- Repeat 3D LC MS/MS analysis of all 8 samples
- A closer examination of the MSMS data (i.e. peptides ID-ed) to understand the cause of the difference in the H:L ratios

Sandia (Protein complexes)

1. Protein Complex work

A. MALDI Analysis

i. A number of bands were identified using anti-HSP70 (E. coli) for pulling down proteins associated with HSP70 during the heat shock process.

ii. These were then subjected to an in-gel tryspsin digest followed by a MALDI-MS (DE PRO Voyager, Applied Biosystems) analysis. Simultaneously the pulled down complex was also directly trypsinized and subjected to a MALDI-MS analysis. The resulting MS spectra were insufficient in resolution to identify the observed bands at the scale they were captured. We are currently working on techniques to enhance MALDI-MS sensitivity.

B. Production of "Tagged" recombinant stress proteins from D vulgaris. In a parallel approach to the antibody-bait-prey capture technique we are working on a direct bait-prey capture technique using recombinantly produced stress response proteins from D. vulgaris. Using homology searches, available microarray data and other resources a list of ~112 proteins involved in a variety of stress responses was provided by Eric Alm. We are currently working on generating appropriate primers for amplifying the corresponding genes. We will be using the GATEWAY Cloning system (Invitrogen) to introduce the amplified genes in a variety of host vectors providing the necessary N- or C-terminal tags.

2. 2D DIGE Work Results

i. Heat shock (50C, 60m): A total of 1414 spots were observed out which 291 displayed differential expression (at least 2 fold) - 92 up-regulated and 206 down-regulated. ii. Oxygen stress (Terry's samples): A total of 1356 spots were observed out of which 126 displayed differential expression (at least 2 fold) - 40 up-regulated and 86 down-regulated. We are currently in the process of analyzing the trypsinized spots through MALDI-MS to get protein ids for spots displaying differential expression.

UCB, LBNL (Metabolomics)

Objective

- Comprehensive analysis of *E. coli* cationic and anionic metabolites by CE-MS

Progress since last report

- CE-MS separation of nucleotides and CoAs (Figure 2 & 3) using a DB-1 column. Application of this method to *E. coli* cell extracts was not successful due to unknown reasons.
- CE-MS separation of amino acids, bases and nucleosides from *E. coli* cell extracts was accomplished. (Figure 4)
- First *D. vulgaris* experiment (oxygen stress responsive) was not successful because I could not get enough cell extracts to analyze from the *D. vulgaris* cell samples.

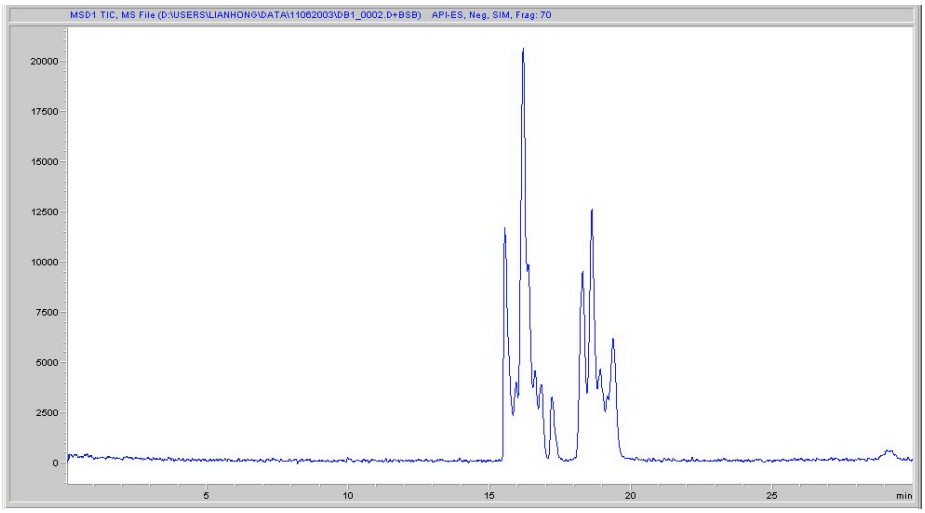


Figure 2. CE-MS analysis of nucleotides standard mixtures. The sample contains: ATP, ADP, AMP, dATP, CTP, CDP, CMP, dCTP, GTP, GDP, GMP, dGTP, TTP, TDP, TMP, UTP, UDP, UMP, dUTP, dAMP, dCMP and IMP. A DB-1 capillary was employed.

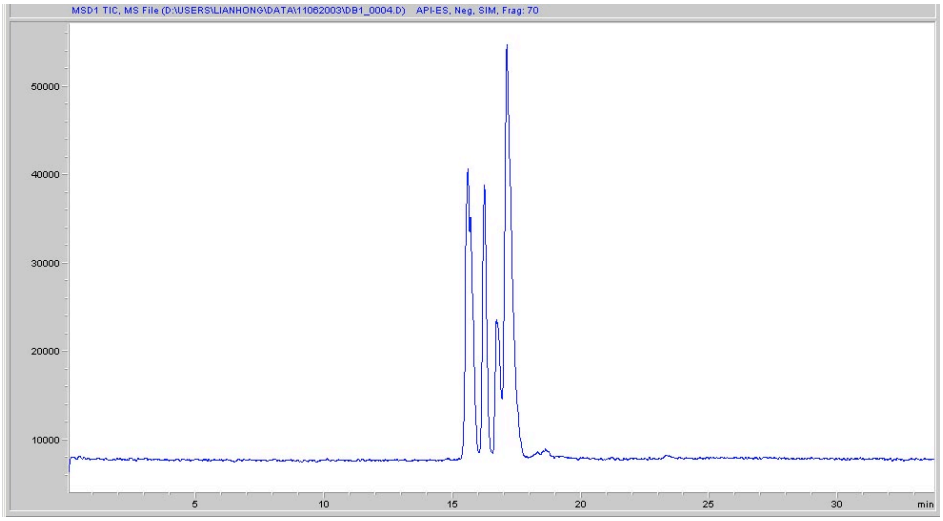


Figure 3. CE-MS analysis of organic acid CoAs standard mixture. The sample contains: CoAs, Acetyl CoAs, malonyl CoAs, Propionyl CoAs and Succinyl CoAs. A DB-1 capillary was employed.

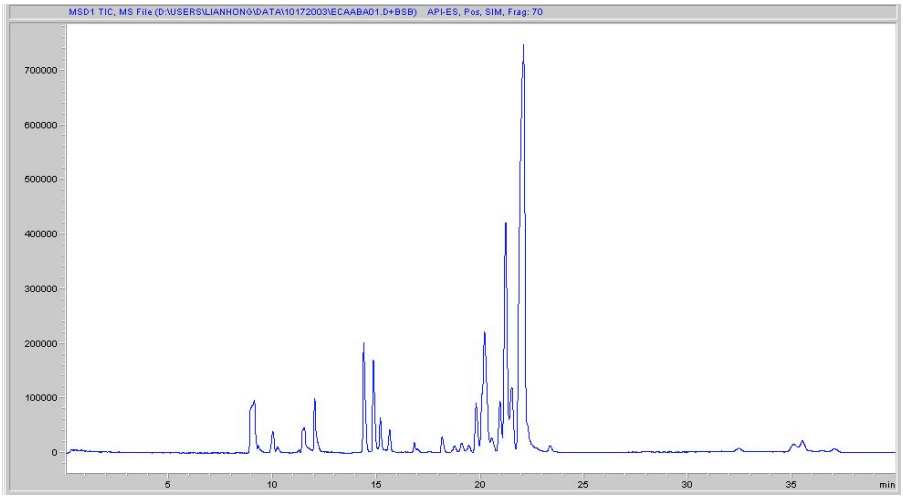


Figure 4. CE-MS analysis of *E. coli* cell extracts. The following metabolites were selected monitored: uridine, cytidine, guanosine, thymidine, adenine, inosine, xanthine, guanine, hypoxanthine, cytosine, adenosine, uracil, 2-deoxycytidine, 2-deoxyadenosine, 2-deoxyuridine, 2-deoxyguanosine, xanthosine, glycine, alanine, valine, leucine, isoleucine, methionine, proline, cysteine, phenylalanine, tyrosine, tryptophan, arginine, lysine, histidine, aspartate, glutamate, serine, threonine, asparagines and glutamine.

Future work

- Analysis of more cationic metabolites from *E. coli* by CE-MS + silica columns
- Analysis of more anionic metabolites from *E. coli* by CE-MS + SMILE columns
- Apply these methods to analyze *D. vulgaris* cell extracts.

University of Missouri, UCB, LBNL (Genetic manipulations)

Objectives

- Develop transposon and directed mutagenesis methods for *D. vulgaris*

A. Selection of transposon mutations

Experiments to select for transposon mutational events require the elimination of the *E. coli* donors. Initial attempts to use differential chloramphenicol concentrations (a concentration to which the wild type SRB is resistant but to which the *E. coli* is sensitive) proved unsatisfactory because of the variability in the cell concentration exposed to the drug. Currently we are attempting to isolate mutants spontaneously resistant to rifampicin or nalidixic acid.

B. Gene transfer into *D. vulgaris*

Electroporation experiments appear to be generating a few recombinants relatively routinely. If we can establish reproducibility, we will provide a protocol in next month's report. Experiments to tag proteins for identifying protein complexes has resulted in the construction of a vector for the production of DnaK tagged with a strep-tag.

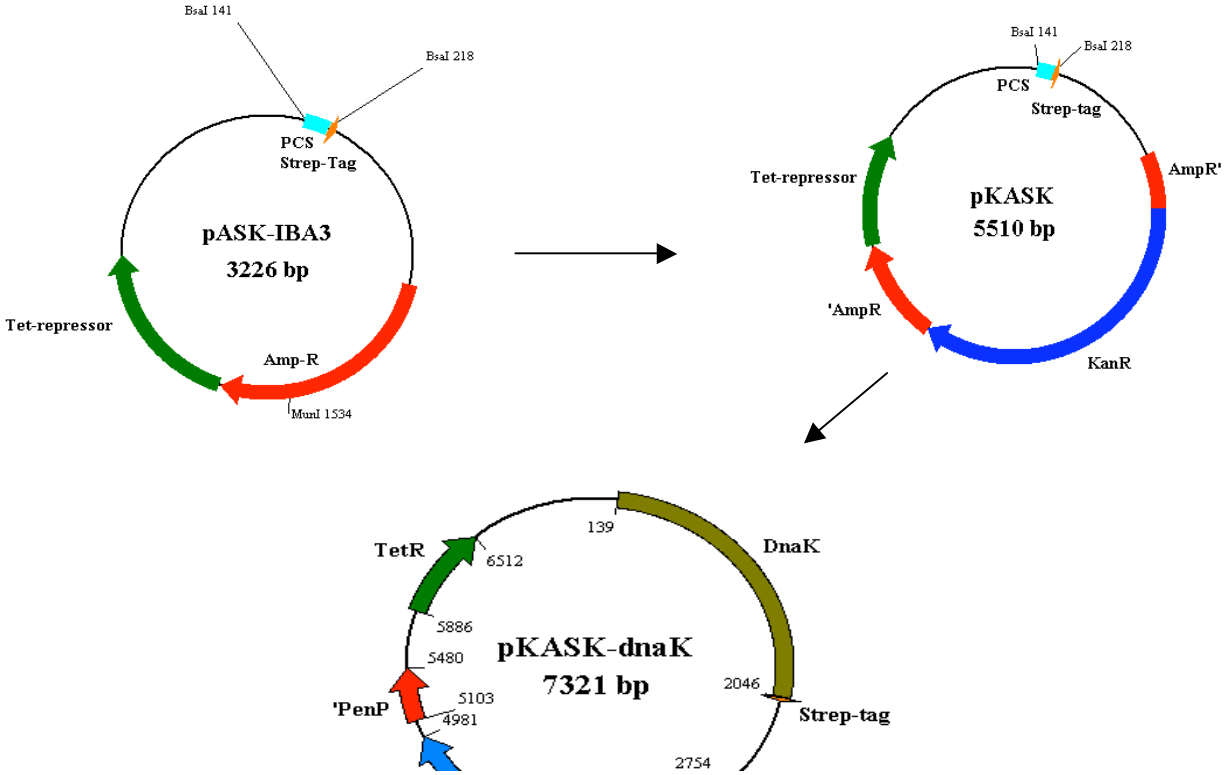
Conjugation: To establish lab protocols to conjugate the knock out strategy vector set (pKOSII, described in the previous report) an MIC for chloramphenicol resistance in *D. vulgaris* was done, using LS4D plates (with KNO_3, 1.5mg/ml). Cam concentrations 0, 10, 20, 30, 40 and 100 ug/ml were checked. *D. vulgaris* was compared with *E.coli* S17-1 with or without the Cam^R plasmid (pKOSII).

In two experiments:
D. vulgaris appears to grow up to Cam 10 plates but not in 20 or higher.
S17-1 pKOSII grows in all plates, Cam0-100.
S17-1 only grows only in Cam 0 plates.
D. vulgaris was also found to be resistant to Gentamycin at 40ug/ml.
To select against S17-1 the conjugation mix will be plated on Cam25Gent40 plates without Nitrate.

C. *Strep*-tagged *dnaK* from *D. vulgaris* in pKASK

The *dnaK* gene (ORF00281) was PCRed with Native *Pfu* polymerase (Invitrogen) from a *D. vulgaris* genomic wizard prep (Promega), ligated into pGEM-T easy (Promega), digested with the *AarI* restriction enzyme (Fermentas) and ligated into the *BsaI* sites of the polycloning site of pKASK. pKASK is a modification of the commercially available pASK-IBA3 plasmid (www.iba-go.com). The modification made to pASK-IBA3 to create pKASK was interruption of the ampicillin resistance cassette (*MfeI* digest, also known as *MunI*; NEB) with the kanamycin cassette from Tn5 (*EcoRI* digest of the pKIXX plasmid).

After *dnaK* had been ligated into pGEM-T easy, sequencing of *dnaK* confirmed the fidelity of the product to the original sequence. Subsequent to construction of pKASK-dnaK, the plasmid was transformed into commercially available a-select cells (silver efficiency, Bioline; www.bioline.com) via heat-shock. Two isolates were screened to ensure the strep-tag region contained an insert of expected size. Both the isolates contained such an insert.



IV. Computational Core

LBNL-Arkin

In October, we added a number of features to the VIMSS comparative genomics web tools in anticipation of making a public release in the next few months. First, we added a comparative microbial genome alignment tool for identifying differences between strains of the same species. In addition, we developed a bug tracking and feature request system to improve response times to user requests. We are near to finishing a prototype "shopping cart" feature that will allow users to store and analyze information about the genes and genomes provided on the VIMSS site, as well as keep track of user identities and login information which will be crucial for our genome annotation plans.

We have scheduled a date of mid-April 2004 with the Joint Genome Institute for the annotation of two sulfate-reducing bacteria: Desulfuromonas acetoxidans and Desulfovibrio desulfuricans (G20). We are working to implement an annotation system that allows users to promote automated annotations to "curated" status, as well as changing and entering new annotations. We are currently working to produce high-quality automated annotations of several related genomes to test our methodology.

In October, we achieved our goal of producing operon predictions for all sequenced bacterial genomes using a completely unsupervised learning algorithm, and have verified it's accuracy on known E. coli and B. subtilis operons. We will continue to work on validating our predictions using microarray data from several species, and submit a manuscript describing the method in December.

We have modified our basic schema for storing gene expression microarray data, and developed some tools for partially automating the data import process. We plan to make these import tools available through the Biofiles website. In addition, we have included gene interactions based on correlation of microarray expression profiles into the regulon browser, and predictions are available on the VIMSS website.

LBNL – Olken

Visual Graph Query User Interface
Vijaya Natarajan joined the project (with funding from the Synechococcus GTL project) at the end of October 2003 to work on the visual graph query user interface (VGQGUI). We have briefed her on the proposed design, given her design documents, etc. Her first task is to select a graph editor toolkit.

We have also evaluated Isaviz, an RDF graph editor. We concluded that the user interface was cumbersome, but that the notion of graph style sheets was very useful and bears adoption by our project. Analogous to HTML/XML style sheets graph style sheets enable specification of graph formatting independently of the graph structure – e.g., choice of icons, arrow shapes, etc. Some limited capability for constructing filtered views of a graph is provided, e.g., to simplify graph

visualization by suppressing some edges (e.g., to water). We have also decided to employ RDF for the configuration files for the VGQGUI.

Graph Data Model
Little has changed on graph data model this month.

Navigational API
This API is being used in the browser development (see below). At this point, only packaging and documentation are needed for completion.

Constructing RDF Query Graphs
We decided to use the Jena toolkit from HP to construct RDF query graphs. Detailed examples, code fragments, etc. have been written and given to Vijaya Natarajan for use in the Visual Graph Query GUI (see above).

Browser
Development of the database schema and instance browser by Kevin Keck continued. This month we added web links to the objects of object properties, i.e., the nodes at the ends of edges in the graph representation of a relational schema or contents.
The browser is nearing completion, but lacks documentation.

Preparing Test Sets for BGDM Initial Demo
This work involves converting B. subtilis dataset of Denise Wolf, et al. from the current DOT file format (used by Graphviz) to RDF. Work began in October and will continue through November.

Graph Algorithm Development
We began work on an algorithm to evaluate partially labeled subgraph homeomorphism queries (i.e., queries which contain path subgraphs specified by regular expressions). Examples include Denise's query containing two back-to-back inhibitory paths. The algorithm being considered is a breadth-first algorithm which should return shorter answers early.

Writing
The WDMBIO workshop report is nearly complete and should ship in the first week of November. An extended abstract concerning this project was submitted to the SIAM Workshop on Scientific Combinatorial Computing, to be held in San Francisco at the end of Feb. 2004. (This at the suggestion of Horst Simon.)

Meetings, Collaborations
Kevin Keck attended the ISWC (International Semantic Web Conference) in Sanibel, Florida. Of particular note was the paper by Robert MacGregor on representing contexts in RDF. MacGregor's proposal is quite similar to that which we have been pursuing.

We hosted the Synechococcus GTL meeting and briefed them on the progress of BGDM. Nagiza Samatova (ORNL) and Ying Xu (ex-ORNL and now Georgia Tech) expressed interest in the use of the system. Both wanted protein interaction networks, in addition to biopathways.

Ying Xu also had a much more general list of data he wanted in the database which resembled BIODB. Finally, there were inquiries about encoding micro-array data as graphs.

Future Work

1) Complete development of RDF/web based relational DB schema/instance browser. (KK)
2) Complete biopathways chapter for DOE Computational Biology Primer. (FO)
3) Complete report on NSF/NLM Workshop on Data Management for Biosciences.(FO)
4) Code DOT file format (from ATT Graphviz) to RDF converter. (KK)
5) Selection of graph toolkit will be used for VGQGUI. (VN, KK, FO)
6) Specification of graph query language RDF encoding. (FO, KK)
7) Continue algorithm design for path queries which satisfy a regular expression. (FO)
8) Contact biopax.org about their efforts on standardization of biopathways data interchange format. (FO,KK)
9) Teach one hour tutorial on biopathways databases on Friday, Nov. 7 for Program in Genomics Applications. (FO)

V. Project Management

- New GTL/VIMSS calendar: The new GTL/VIMSS calendar is online http://vimss.lbl.gov/vimss_calendar.html. All of the GTL meetings will be listed on this website (internal and external announcements). Send additional meeting notifications to Nancy at naslater@lbl.gov.

- The FY04 budgets were sent to all of the PIs in October; however, we are operating under the DOE Continuing Resolution until further notice.

- The deadline for submitting abstracts to the ASM meeting is early December. We would like to have a strong showing for our project this year. Each PI is encouraged to work with other groups within our GTL project on submitting an abstract.

- FY04 Milestones from PIs are due. If you have not already done so, please submit your milestones to Nancy immediately.